doi：10. 12068/j. issn. 1005－3026. 2026. 20260035

基于多层次特征协同建模的抓取姿态估计网络

王华斌，任 涛，杨渊铭
（东北大学 软件学院，辽宁 沈阳 110169）

摘 要：针对抓取姿态估计中全局语义建模不足、局部几何表达受限以及复杂场景下多尺度特征难以统一表征的问题，本文围绕多层次特征协同建模开展研究，并在二维图像与三维点云两类感知表征上构建了对应的网络方法．针对二维平面抓取任务，提出并联Transformer与ResUnet++的PTUNet，通过全局语义分支与局部细节分支的双向交互，实现抓取质量、角度与宽度的联合预测．针对六自由度抓取任务，提出CascadeGraspNet，通过级联局部−全局优化模块和多尺度圆柱体采样模块，增强点云局部结构感知、全局上下文建模以及跨尺度几何聚合能力．实验结果表明，PTUNet在Jacquard数据集上的抓取准确率达到97. 8%；CascadeGraspNet在GraspNet−1Billion数据集RealSense子集上对已知物体、相似物体和新物体的平均精度分别达到71. 39%，63. 29%和28. 69%，并在真实场景抓取实验中取得97. 8%的抓取成功率．由此可见，多层次特征协同建模能够有效提升抓取姿态估计的精度、鲁棒性和泛化能力．

关 键 词：抓取姿态估计；多层次特征协同建模；深度学习；Transformer；点云；六自由度抓取
中图分类号：TP 183 **文献标志码**：A **文章编号**：1005−3026(2026)07−0115−10

Network for Grasp Pose Estimation Based on Multi-level Feature Collaborative Modeling

WANG Hua-bin，*REN Tao*，*YANG Yuan-ming*
（School of Software，Northeastern University，Shenyang 110169，China. Corresponding author：REN Tao，E−mail：chinarentao@163.com）

Abstract：To address the problems of insufficient global semantic modeling，limited local geometric representation，and difficulty in unified representation of multi-scale features in complex scenes in grasp pose estimation，this paper conducted research centered on multi-level feature collaborative modeling and constructed corresponding network methods on two perception representations：two-dimensional images and three-dimensional point clouds. For two-dimensional planar grasp tasks，PTUNet with parallel Transformer and ResUnet++ was proposed. Through the bidirectional interaction between global semantic branch and local detail branch，the joint prediction of grasp quality，angle，and width was realized. For six-degree-of-freedom grasp tasks，CascadeGraspNet was proposed. Through the cascading local-global optimization module and multi-scale cylinder sampling module，the abilities of local structure perception，global context modeling，and cross-scale geometric aggregation of point clouds were enhanced. Experimental results indicate that PTUNet grasp accuracy reaches 97. 8% on the Jacquard dataset；CascadeGraspNet average precision for known objects，similar objects，and novel objects reaches 71. 39%，63. 29%，and 28. 69%，respectively，on the RealSense subset of GraspNet−1Billion dataset，and it achieves the grasp success rate of 97. 8% in real-scene grasp experiments. Therefore，multi-level feature collaborative modeling can effectively improve the accuracy，robustness，and generalization capability of grasp pose estimation.

Key words：grasp pose estimation；multi-level feature collaborative modeling；deep learning；Transformer；point cloud；six-degree-of-freedom grasp

收稿日期：2026−03−25
基金项目：国家自然科学基金资助项目(62276058)；中央高校基本科研业务费专项资金资助项目(N25GFZ011).
作者简介：王华斌(1989—)，男，河南确山人，东北大学博士研究生.
通信作者：任 涛，E−mail：chinarentao@163.com.

抓取姿态估计是智能操作系统中的关键感知与决策环节,其目标是在复杂视觉输入中确定稳定、可执行且具备几何可行性的抓取位姿.随着深度学习方法在图像理解、三维场景分析与端到端控制中的持续发展,抓取姿态估计逐步从依赖人工规则和手工特征的传统范式,转向依赖数据驱动表征学习的深度网络范式[1].相比仅强调具体执行平台的系统性工作,当前相关研究的竞争核心已更多地集中在网络对语义信息、结构信息以及跨尺度特征的综合建模能力上.

现有方法可分为面向二维图像的平面抓取估计和面向点云的六自由度抓取估计两类.前者通常直接从RGB(red green blue)或RGB-D(red green blue-depth)图像中预测抓取位置、角度和夹爪宽度[2-4],具有计算效率高、部署成本低等优点,但传统卷积结构受限于局部感受野,难以稳定刻画目标整体形状、主轴方向及其与背景之间的长距离依赖关系;后者能够在三维空间中表达更灵活的抓取位姿,更适合处理遮挡、堆叠和非结构化场景[5-8],但局部几何细节提取、全局上下文建模以及多尺度邻域建模之间往往难以兼顾.

从表征学习角度看,抓取姿态估计中的关键问题可归结为特征建模层次不足:缺乏全局语义会影响抓取方向与物体整体结构的一致性,缺少局部细节则易导致精确定位、夹爪开口估计和避碰判断误差,而单一尺度建模也难以适应复杂场景中的尺度变化.基于此,本文以多层次特征协同建模为统一主线,将二维抓取与六自由度抓取置于同一研究框架下,强调在网络中协同融合全局与局部、语义与几何、单尺度与多尺度信息,并通过显式交互机制实现特征对齐、补偿与融合.围绕这一思路,本文分别在图像域和点云域设计了PTUNet与CascadeGraspNet:PTUNet采用Swin Transformer与ResUnet++并联双分支结构,并结合双向特征交互和双注意力增强机制,实现全局语义与局部细节的互补建模;CascadeGraspNet则通过Cascade-LGR(Cascade local-global refinement)模块强化局部邻域与全局上下文的级联优化,并借助多尺度圆柱体采样(multi-scale cylinder grouping,MCG)模块完成与夹爪接近方向对齐的跨尺度几何聚合,从而提升复杂且杂乱场景下抓取候选点精细结构与物体整体上下文的联合表达能力.

本文的主要贡献如下:1)提出以多层次特征协同建模为核心的统一分析思路,从全局与局部、语义与几何、单尺度与多尺度三个层面阐释抓取姿态估计中的关键建模问题;2)针对二维平面抓取任务,提出PTUNet,通过以Transformer[9]为基础的Swin Transformer[10]分支与ResUnet++[11]分支以及双向特征交互机制,实现全局语义与局部细节的协同建模;3)针对六自由度抓取任务,提出CascadeGraspNet,通过Cascade-LGR模块与MCG模块实现局部几何、全局上下文和跨尺度结构的协同表达;4)在公开数据集与真实场景实验中验证了所提方法的有效性,结果表明所提方法在抓取精度、鲁棒性和泛化能力方面均具有较好表现.

1 方 法

1.1 多层次特征协同建模总体思路

本文将抓取姿态估计问题抽象为任务相关特征的层次化表达与协同解码问题.在这一视角下,网络设计不应只关注某一层特征的增强,而应关注不同层级特征之间是否形成互补关系.对于图像输入,需要在全局语义和局部细节之间建立高效的信息流,使模型既能理解目标与背景的整体关系,又能保留抓取中心、边缘轮廓和局部纹理等精细信息;对于点云输入,则需要在局部邻域几何与全局空间结构之间建立稳定联系,并同时处理尺度变化带来的表征偏差.围绕上述目标,本文提出的多层次特征协同建模框架包含三个基本原则:第一,采用双路径或多阶段结构分别承担不同层级特征的提取任务,避免单一路径过度折中;第二,通过显式交互模块完成异质特征的对齐与互补,而非仅在末端进行简单拼接;第三,将网络输出设计为与任务几何意义一致的多分支或多任务形式,使不同参数在统一监督下共同优化.基于这三个原则,PTUNet和CascadeGraspNet在结构形式上虽不相同,但都体现出先分层提取、再协同融合、最后联合预测的基本思路.

1.2 面向二维图像的全局-局部协同建模网络

1.2.1 抓取表示

对于二维平面抓取任务,本文采用基于像素级热力图的抓取表示方式.相较于直接回归五维抓取矩形参数的方法,热力图表示可以更自然地保留图像的空间布局,并将抓取姿态估计分解为抓取质量、抓取角度和抓取宽度3个预测分支.设Q,A,W分别表示抓取质量热力图、角度热力图和宽度热力图,网络可通过对质量热力图中高响应区域进行搜索,并结合对应位置的角度和宽度预测结

果生成最终抓取姿态.该方式有利于多任务联合优化,并能减少角度周期性带来的回归不稳定问题.

1.2.2　PTUNet网络结构

图1展示了PTUNet的网络结构.该网络采用由ST(Swin Transformer)分支和RU(ResUnet)分支组成的并联双分支架构,分别用于全局语义建模和局部细节提取.输入RGB-D图像同时送入2条分支,并通过跨分支特征交互实现信息互补,最终联合预测抓取质量、抓取角度和抓取宽度热力图.ST分支基于分层式Swin Transformer编码器-解码器结构,通过窗口多头自注意力和移动窗口多头自注意力建模目标整体形状、主轴方向及长距离依赖关系,为抓取姿态预测提供全局语义先验信息;RU分支基于ResUnet++构建,利用残差卷积块、多级跳跃连接和通道重标定机制提取边缘、纹理和轮廓等细粒度特征,为抓取中心定位和局部结构恢复提供高分辨率空间信息.通过双分支协同设计,PTUNet实现了全局语义与局部细节的互补建模,为抓取参数联合预测提供了更充分的特征支撑.

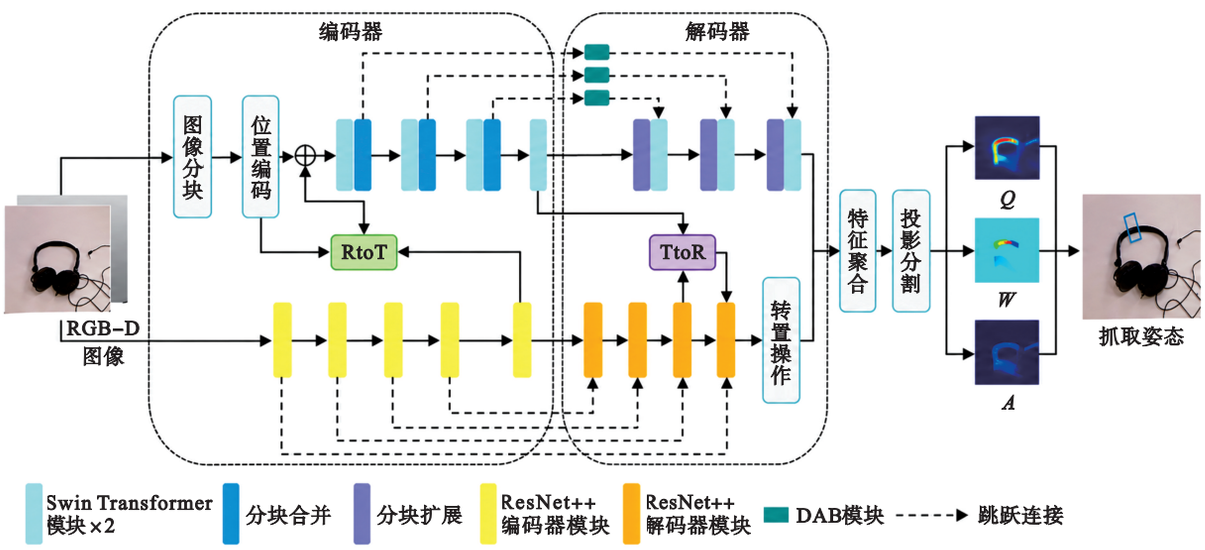


图1　PTUNet网络结构

Fig. 1　PTUNet network structure

1.2.3　双向特征交互与注意力增强

为了避免双分支仅形成并行堆叠而缺乏有效协同,PTUNet设计了特征交互模块(feature interaction module,FIM),在编码与解码阶段分别引入双向特征交互模块.图2展示了RtoT(ResUNet to Transformer)模块的结构,该模块在编码阶段将RU分支提取的高层局部语义特征映射到Transformer嵌入空间,使ST分支在全局建模过程中能够保留更多边缘和结构细节信息.

图3展示了TtoR(Transformer to ResUNet)模块的结构,该模块在解码阶段将ST分支中经全局建模后的上下文特征回注到RU分支,从而增强局部特征与整体语义之间的一致性.

除双向交互外,本文还在ST分支的跳跃连接中引入双注意力模块(dual attention block,DAB),其结构如图4所示.

该模块由通道注意力(channel attention module,CAM)和位置注意力(position attention module,PAM)两部分组成.

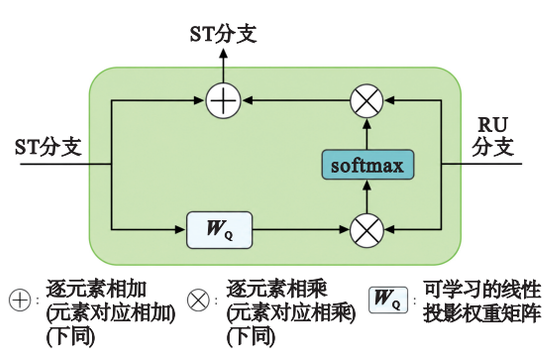


图2　RtoT模块结构

Fig. 2　RtoT module structure

通道注意力通过全局统计信息自适应调整各通道响应强度,以突出与抓取相关的语义特征;位置注意力通过建模空间位置之间的相关性,强化抓取中心及其邻域区域的响应.两者联合后,能够在保留多尺度信息的同时抑制背景噪声和无关纹理的干扰.通过特征交互模块与双注意力模块的协同作用,PTUNet不仅增强了双分支

之间的信息流动,也提升了关键区域的判别能力,从而提高了抓取姿态预测的稳定性和一致性.

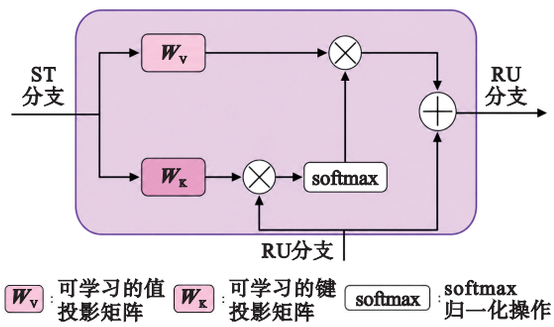


图3 TtoR模块结构

Fig. 3 TtoR module structure

1.2.4 多任务损失设计

为实现抓取质量、抓取角度和抓取宽度的联合优化,PTUNet采用基于Huber损失[12]的多任务损失函数进行训练.总损失由抓取质量分支损失、抓取角度分支损失和抓取宽度分支损失加权求和构成,即 $L=\lambda_1 L_Q+\lambda_2 L_A+\lambda_3 L_W$. 式中:$L$ 为总损失函数;L_Q,L_A 和 L_W 分别为抓取质量、抓取角度和抓取宽度的预测损失;λ_1,λ_2 和 λ_3 为各分支的权重系数.与单纯采用均方误差相比,Huber损失在误差较小时具有平方误差的平滑优化特性,在误差较大时又能减弱异常值对训练过程的影响,因此更适合抓取姿态估计中的多参数联合回归任务.对于角度分支,本文结合角度分解编码方式进行监督,以缓解角度周期性带来的不连续问题.通过上述联合损失设计,网络能够在统一优化框架下兼顾抓取中心定位、方向估计和夹爪宽度预测,从而提升抓取参数之间的一致性与整体预测稳定性.

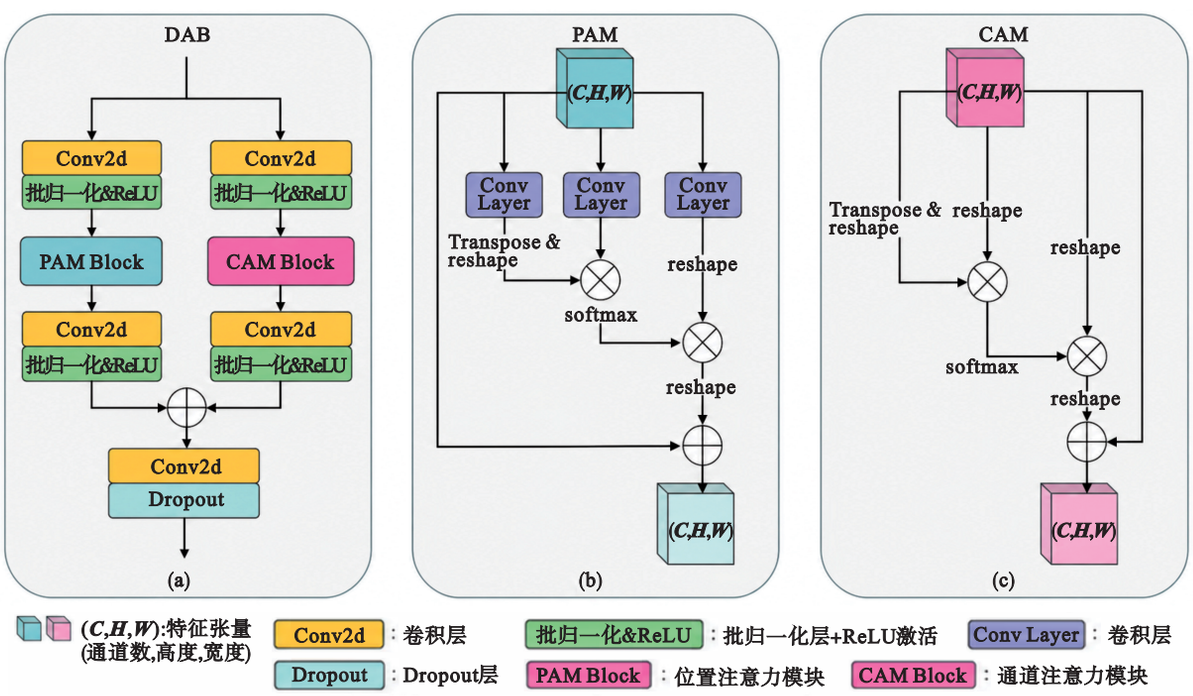


图4 DAB模块结构

Fig. 4 DAB module structure

1.3 面向三维点云的局部-全局协同建模网络

1.3.1 六自由度抓取表示

对于六自由度抓取任务,直接回归完整的旋转矩阵往往优化难度较高.为此,本文采用分解式姿态表示,将抓取位姿分解为抓取中心位置、接近方向、平面内旋转角和夹爪宽度与深度等参数.网络首先通过视点分类确定夹爪接近方向,再对平面内旋转角和夹爪宽度进行回归,并联合估计夹爪开口宽度与抓取质量.该表示方式既降低了旋转学习的复杂度,也更符合抓取执行时的物理约束.

1.3.2 CascadeGraspNet总体结构

图5展示了CascadeGraspNet的网络结构.其中,N 表示输入点云的点数;f 表示网络层数;C 表示点特征通道数;S 表示经可抓取性筛选和最远点采样获得的种子点数;V 表示离散候选接近方向数;D_v 表示视图网络输出的视图嵌入特征数.该网络以稀疏点云为输入,整体由特征提取、候选抓取点生成、视点选择、局部几何建模和抓取参数预测5个部分组成.

与直接从全局点云一次性回归抓取姿态的方法不同,CascadeGraspNet采用分阶段建模思

路,先筛选具有抓取潜力的候选区域,再围绕候选点开展精细几何建模,从而在控制计算复杂度的同时提升姿态估计精度.网络主干采用基于 MinkowskiEngine[13]构建的 LGR-SparseUNet.该主干网络遵循编码器-解码器结构,在逐级下采样过程中提取点云的多层次语义特征,并在编码器各级残差块后嵌入 Cascade-LGR 模块,以持续增强点云特征表达能力.解码阶段通过逐级上采样和跳跃连接恢复空间分辨率,并融合深浅层信息,为后续候选点判别和参数预测提供丰富的上下文表示.在候选抓取点生成阶段,编码后的点云特征首先输入 GraspableNet,完成抓取可行性判别并筛选高置信度候选点;随后通过 ViewNet 对候选点的接近方向进行预测,为后续局部几何建模提供方向先验信息.最后,围绕候选点及其接近方向构建局部邻域,并在此基础上完成抓取参数回归.上述结构使网络能够同时兼顾全局场景理解与局部几何分析,为复杂场景下的六自由度抓取姿态估计提供了有效的分层建模框架.

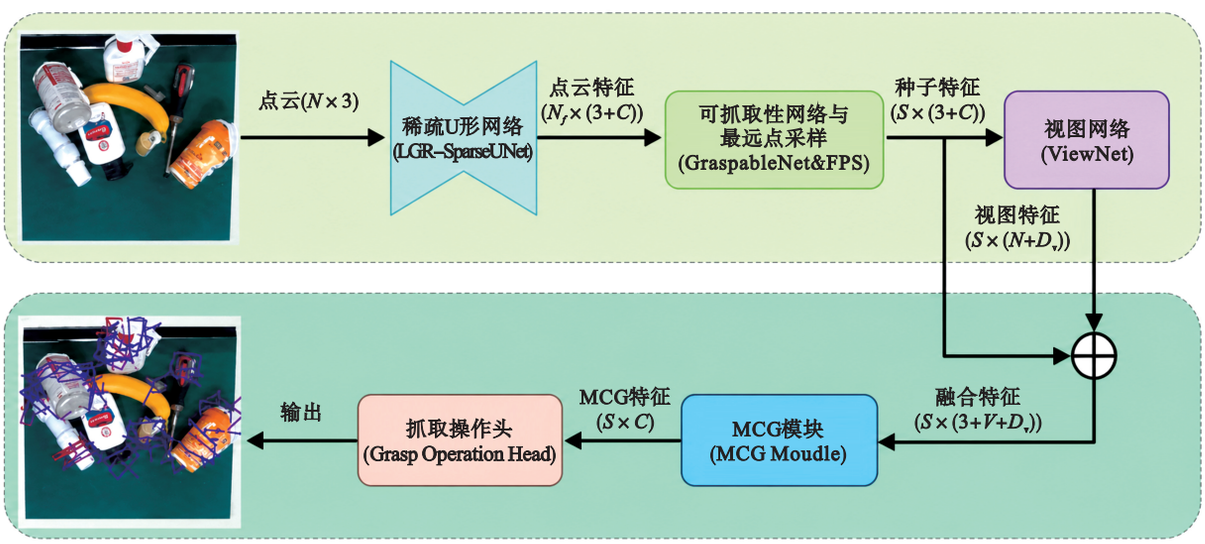


图 5　CascadeGraspNet 网络结构

Fig. 5　CascadeGraspNet network structure

1.3.3　Cascade-LGR 模块

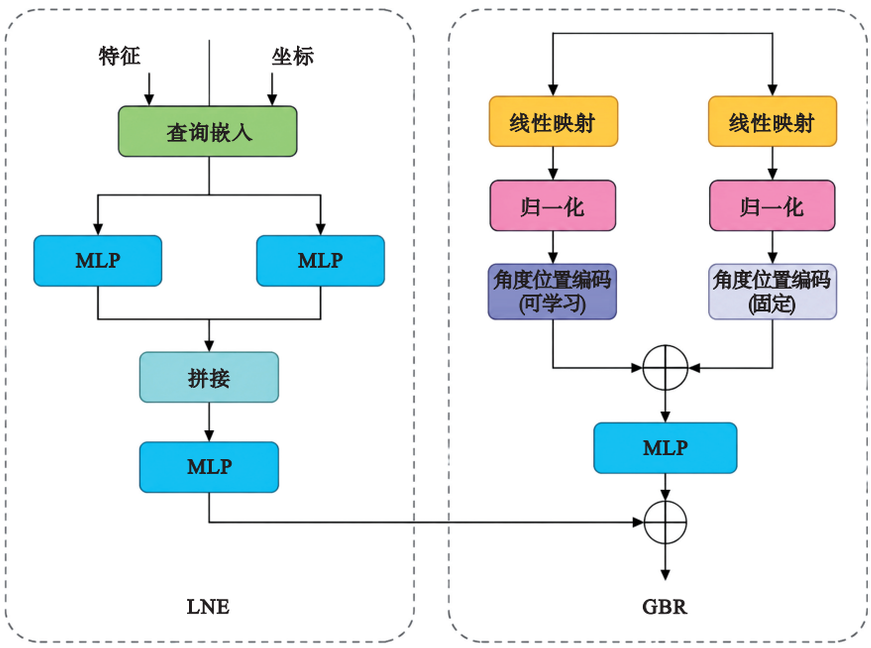
图 6 为 Cascade-LGR 模块的结构.该模块由局部邻域特征增强(local neighborhood enhancement,LNE)和全局双线性正则化(global bilinear regularization,GBR)两个阶段组成,其目标是在同一特征提取单元中实现局部几何信息与全局上下文信息的级联优化.

在 LNE 阶段,模块围绕每个候选点构建动态邻域,通过 K 近邻搜索获取局部点集,并同时编码邻域内的几何偏移与特征差异,形成更具判别性的局部上下文表示.该阶段重点增强点级特征对边缘结构、表面法向以及局部接触区域的感知能力,从而提高网络对可抓取区域的识别精度.在 GBR 阶段,模块进一步从通道维和空间维建模长程依赖关系,通过构造全局响应矩阵对局部增强特征进行重整合与约束,以抑制背景区域的冗余响应并强化与抓取相关的全局上下文模式.经过两阶段级联优化后,模型不仅能够关注抓取点附近的局部结构,还能够兼顾目标整体形状及周围环境带来的全局约束.因此,Cascade-LGR 模块实质上建立了从局部几何到全局语义的连续优化路径,为后续抓取姿态预测提供了更加稳定和一致的点云特征表示.

1.3.4　多尺度圆柱体采样与跨尺度融合

在候选抓取点及接近方向确定后,CascadeGraspNet 通过多尺度圆柱体采样(MCG)模块进行局部几何建模(R 为圆柱邻域半径,r=0.02 m,为基础半径),如图 7 所示.MCG 沿夹爪接近方向,在候选点周围构建半径为 0.02,0.04,0.06 m 的多尺度圆柱邻域,使采样区域更符合实际抓取空间约束,并覆盖从局部细节到夹爪尺度结构的不同层次.各尺度下,模块采用最远点采样选取代表点,并结合点坐标与法线信息提取几何特征.其中,小尺度邻域侧重边缘、凹槽和接触面等细粒度结构,大尺度邻域则用于刻画目标整体布局及其与周围物体的空间关系.随后,MCG 将多尺度特征在通道维拼接,并通过自注意力机制建立跨尺度关联,实现局部细节与整体结构的

动态融合.该模块将夹爪接近方向先验信息融入局部特征提取过程,可缓解固定感受野带来的尺度偏差,提升模型在小目标、遮挡和密集堆叠场景下的几何建模能力.



注:MLP为多层感知器(multi-layer perceptron).

图6 Cascade-LGR模块结构

Fig. 6 Cascade-LGR module structure

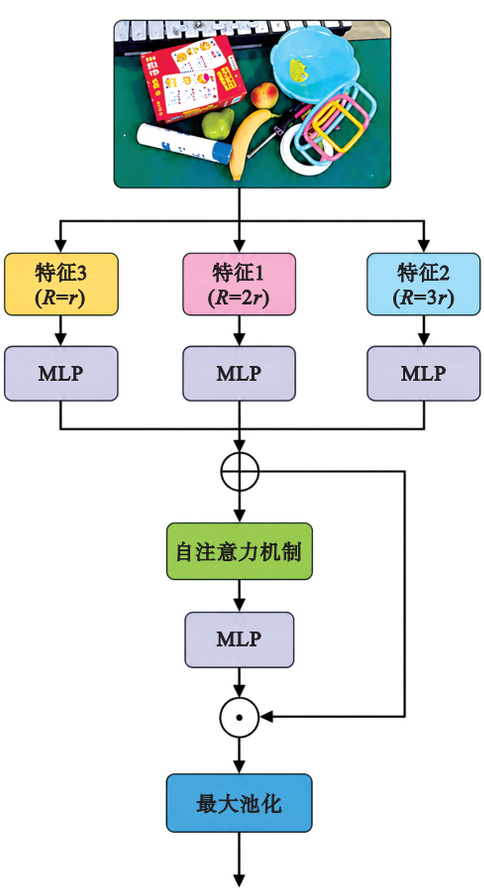


图7 MCG模块结构

Fig. 7 MCG module structure

1.3.5 多任务联合优化

为提升六自由度抓取姿态估计的准确性,CascadeGraspNet采用多任务联合损失进行训练,对抓取可行性、抓取中心位置、视角方向、夹爪宽度、旋转类别、抓取评分以及深度一致性进行统一监督.总损失函数(L_{total})定义为各子任务损失的加权线性组合,如式(1)所示:

$$L_{\text{total}}=\alpha_1 L_{\text{cls}}+\alpha_2 L_{\text{center}}+\alpha_3 L_{\text{view}}+\alpha_4 L_{\text{width}}+\alpha_5 L_{\text{rot}}+\alpha_6 L_{\text{score}}+\alpha_7 L_{\text{depth}}. \tag{1}$$

其中:L_{cls}为抓取可行性分类损失(二元交叉熵),权重α_1=1.0;L_{center}为抓取中心位置回归损失(平滑$L1$损失),权重α_2=2.0;L_{view}为视角方向预测损失(交叉熵),权重α_3=1.0;L_{width}为夹爪宽度回归损失(平滑$L1$损失),权重α_4=1.0;L_{rot}为旋转类别分类损失(交叉熵),权重α_5=1.0;L_{score}为抓取评分回归损失(平滑$L1$损失),权重α_6=1.0;L_{depth}为深度一致性约束损失(平滑$L1$损失),权重α_7=0.5.抓取中心位置损失权重适当加大(α_2=2.0),以强化对精确抓取中心定位的监督;深度一致性损失权重设为0.5,在引入几何约束的同时避免对主要任务的过度干扰.通过对抓取可行区域判别、姿态参数预测和几何一致性约束的联合建模,网络能够在同一优化框架下同时学习抓取位置、接近方向

和操作参数,从而提升复杂场景下抓取姿态估计的准确性、稳定性和物理合理性.

2　实　　验

2.1　数据集与实验设置

为验证所提方法在不同感知模态和不同抓取任务中的有效性,本文分别在Jacquard和GraspNet-1Billion数据集上开展实验.Jacquard数据集[14]基于ShapeNet[15]模型与物理引擎构建,包括54 485张RGB-D图像和大规模仿真抓取标注,适合检验模型在大规模多样化样本上的学习能力.GraspNet-1Billion[16]则是当前六自由度抓取检测领域最具代表性的基准之一,提供多视角RGB-D采集结果、密集点云以及超过11亿个抓取姿态标注.

在二维抓取实验中,采用Jacquard标准划分方式进行训练和评估,评价指标以抓取矩形的Jaccard指数和角度误差阈值为准;在六自由度抓取实验中,采用平均精度对已知物体、相似物体和新物体三类场景进行评估,并在评测前执行非极大值抑制以避免相近姿态的重复统计.

2.2　二维抓取实验结果分析

表1给出了本文方法与其他方法在Jacquard数据集上的对比结果.

表1　在Jacquard数据集上的对比实验结果

Table 1　Comparative experimental results on Jacquard dataset

方法	输入	准确率/%	参数量/M	耗时/(ms·帧$^{-1}$)	帧率/(帧·s^{-1})
Jacquard [14]	RGB-D	74.2	—	—	—
GG-CNN2 [17]	D	84.0	0.30	14.2	70.4
FCGN [18]	RGB	91.8	6.97	22.1	45.2
GQ-STN [19]	D	70.8	11.30	35.6	28.1
ROI-GD [20]	RGB	90.4	8.50	28.4	35.2
DetSegRefine [21]	RGB	93.0	12.10	31.8	31.4
GR-ConvNet [22]	RGB-D	94.6	1.90	18.3	54.6
TF-GraspNet [23]	RGB-D	94.6	25.60	48.2	20.7
DSC-GraspNet [24]	RGB-D	94.7	9.80	26.5	37.7
PTUNet	RGB-D	97.8	38.20	32.6	30.7

注:“—”表示原始文献未报道对应指标;耗时在NVIDIA RTX 3090(batch size=1)条件下测量.

PTUNet的抓取准确率达到97.8%,表明所提方法能够有效提升二维抓取姿态估计性能.该结果说明,并联双分支结构能够较好地兼顾全局上下文建模与局部细节保持,其中Transformer分支有助于捕获目标整体结构及其与背景之间的长距离依赖关系,ResUnet++分支则有助于增强边缘、纹理和局部轮廓等细粒度特征表达,从而提高抓取中心定位与姿态预测的一致性.

为进一步评估本文方法的工程适用性,表1给出了各方法的模型参数量、耗时及帧率指标.尽管Transformer分支引入了全局注意力计算,但本文通过合理设计窗口大小与分支融合策略,将推理开销控制在可接受范围内,相较于同精度的TF-GraspNet具有明显的效率优势.

表2给出了PTUNet在Jacquard数据集上的消融实验结果.

基线模型的准确率为87.8%,引入RU Branch后提升至89.2%;在此基础上加入特征交互模块(FIM)后,准确率提升至91.2%;加入双注意力模块(DAB)后,准确率达到95.7%;同时引入FIM与DAB后,最终准确率提升至97.8%.实验结果表明,FIM能够有效增强全局语义与局部细节之间的对齐与融合,DAB则进一步提升了模型对关键抓取区域的响应能力,两者结合后能够显著提高模型整体性能.

表2　在Jacquard数据集上的消融实验结果

Table 2　Ablation experimental results on Jacquard dataset

Baseline	RU Branch	FIM	DAB	准确率/%
√				87.8
√	√			89.2
√	√	√		91.2
√	√		√	95.7
√	√	√	√	97.8

2.3　六自由度抓取实验结果分析

表3给出了本文方法与其他方法在GraspNet-1Billion数据集RealSense子集上的对比结果.CascadeGraspNet在已知物体、相似物体和新物体三类场景下的平均精度(average precision,AP)分别达到71.39%,63.29%和28.69%,均表现出较好的性能.与TransGrasp[25],TSBGrasp[26],HGGD[27],GSNet[28],EconomicGrasp[29]和E3GNet[30]等方法相比,所提方法在已知物体和相似物体场景中具有更强的建模能力,在新物体场景中也保持了较好的泛化性能,说明该网络对真实传感器采集的点云具有较强适应性.其性能提升主要得益于Cascade-LGR对局部几何与全局上下文的

级联优化,以及MCG对多尺度结构信息的有效聚合.此外,由于GraspNet-1Billion基准上部分近期方法仅在Kinect传感器子集上公开结果,本文统一采用RealSense子集进行评测,以保证对比的公平性.在该子集的公开对比结果中,本文方法在已知物体场景下的AP达到71.39%,处于领先水平,进一步验证了所提方法的有效性与竞争力.

表3中AP0.8和AP0.4分别表示在摩擦系数为0.8和0.4条件下计算的平均精度.CascadeGraspNet在精度显著领先的同时,推理速度与EconomicGrasp和GSNet等轻量化方法处于同一量级.基于MinkowskiEngine稀疏卷积的LGR-SparseUNet骨干网络,其计算复杂度与实际有效体素数量成线性关系,在典型杂乱场景(约20 000点)下相较于密集卷积方案具有显著的计算优势.MCG模块的三尺度自注意力融合复杂度为$O(S\cdot n_s^2)$(S=3,n_s≤64).

表3 在GraspNet-1Billion数据集RealSense子集上的对比实验结果

Table 3 Comparative experimental results on RealSense subset of GraspNet-1Billion dataset %

方法	已知物体			相似物体			新物体		
	AP	AP0.8	AP0.4	AP	AP0.8	AP0.4	AP	AP0.8	AP0.4
GPD	22.87	28.53	12.84	21.33	27.83	9.64	8.24	8.89	2.67
PointnetGPD	25.96	33.01	15.37	22.68	29.15	10.76	9.23	9.89	2.74
GraspNet	27.56	33.43	16.95	26.11	34.18	14.23	10.55	11.25	3.98
TransGrasp[25]	39.81	47.54	36.42	29.32	34.80	25.19	13.83	17.11	7.67
TSBGrasp[26]	58.95	68.18	54.88	52.97	63.24	46.99	22.63	28.53	12.00
HGGD[27]	58.35	66.54	55.96	47.93	56.91	41.86	22.10	27.37	14.31
GSNet[28]	65.70	76.25	61.08	53.75	65.04	45.97	23.98	29.93	14.05
EconomicGrasp[29]	68.21	**79.60**	63.54	61.19	**73.60**	53.77	25.48	31.46	13.85
E3GNet[30]	69.56	78.35	66.30	60.84	72.81	52.07	26.74	33.12	15.43
Ours	**71.39**	78.21	**68.55**	**63.29**	73.08	**55.86**	**28.69**	**35.73**	**16.96**

注:黑体表示实验结果最佳.

表4给出了本文方法在GraspNet-1Billion数据集RealSense子集上的消融实验结果.

表4 在GraspNet-1Billion数据集RealSense子集上的消融实验结果

Table 4 Ablation experimental results on RealSense subset of GraspNet-1Billion dataset %

Baseline	Cascade-LGR	MCG	已知物体			相似物体			新物体		
			AP	AP0.8	AP0.4	AP	AP0.8	AP0.4	AP	AP0.8	AP0.4
√			65.70	76.25	61.08	53.75	65.04	45.97	23.98	29.93	14.05
√	√		70.31	78.35	67.77	61.24	71.38	54.37	26.10	34.16	16.43
√		√	66.12	77.82	65.36	60.13	69.12	52.24	24.75	31.83	15.49
√	√	√	71.39	78.21	68.55	63.29	73.08	55.86	28.69	35.73	16.96

基线模型在已知物体类别上的平均精度为65.70%,仅加入Cascade-LGR时提升至70.31%,仅加入MCG时提升至66.12%,同时引入2个模块后达到71.39%.这表明Cascade-LGR与MCG均能够带来稳定增益,其中Cascade-LGR对当前实验设置下的性能提升更为明显,而两者联合使用时能够进一步增强模型对复杂场景中局部结构、全局关系和尺度变化的综合建模能力.

如表5所示,三尺度组合{0.02 m,0.04 m,0.06 m}相较于任意单尺度配置均取得显著性提升($AP_{已知}$提升1.6~3.0个百分点),验证了多尺度圆柱邻域设计的必要性.进一步增加至四尺度时,性能出现轻微下降,且计算开销增加约30%,这表明三尺度配置在性能与效率之间取得了最优平衡.

表5 MCG模块圆柱邻域半径组合消融实验(GraspNet-1Billion RealSense子集)

Table 5 Ablation experiment on cylinder neighborhood radius combinations of MCG module (GraspNet-1Billion RealSense subset)

半径组合/m	尺度数	$AP_{已知}$/%	$AP_{相似}$/%	$AP_{新物体}$/%
{0.02}	单尺度(小)	68.43	60.12	25.87
{0.04}	单尺度(中)	69.77	61.34	26.92
{0.06}	单尺度(大)	69.21	60.98	26.41
{0.02,0.04}	双尺度	70.32	62.17	27.63
{0.02,0.06}	双尺度(跨度大)	70.81	62.54	28.01
{0.02,0.04,0.06}	三尺度	71.39	63.29	28.69
{0.01,0.02,0.04,0.06}	四尺度	71.12	63.05	28.42

3　结　　语

本文针对抓取姿态估计中的特征建模问题,提出了以多层次特征协同建模为核心的统一研究思路,并面向二维图像与三维点云两类感知表征分别构建了 PTUNet 和 CascadeGraspNet 两种网络方法:前者通过并联 Swin Transformer 与 ResUnet++结构,并引入双向特征交互机制,实现全局语义信息与局部细节特征的协同建模;后者则借助 Cascade-LGR 模块与 MCG 模块,增强局部几何、全局上下文及跨尺度结构之间的协同表达.实验结果表明,所提方法在二维抓取和六自由度抓取任务中均取得了较好性能,其中 PTUNet 在 Jacquard 数据集上的抓取准确率达到 97.8%, CascadeGraspNet 在 GraspNet-1Billion 数据集 RealSense 子集上针对已知物体、相似物体和新物体分别取得 71.39%, 63.29% 和 28.69% 的平均精度,并在真实场景抓取实验中获得 97.8% 的抓取成功率,验证了通过结构化方式实现不同层次特征协同建模对于提升抓取姿态估计精度、鲁棒性和泛化能力的有效性.未来可进一步加强图像语义与点云几何之间的跨模态融合,统一不同感知表征下的协同建模机制,同时优化网络结构与推理效率以适应实时部署和边缘计算需求,并结合自监督学习与强化学习策略,提升模型在开放环境中的持续适应能力.

参考文献:

[1] Bohg J, Morales A, Asfour T, et al. Data-driven grasp synthesis: a survey[J]. *IEEE Transactions on Robotics*, 2014, 30(2): 289-309.

[2] Jiang Y, Moseson S, Saxena A. Efficient grasping from RGBD images: learning using a new rectangle representation[C]//2011 IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE, 2011: 3304-3311.

[3] Lenz I, Lee H, Saxena A. Deep learning for detecting robotic grasps[J]. *The International Journal of Robotics Research*, 2015, 34(4/5): 705-724.

[4] Redmon J, Angelova A. Real-time grasp detection using convolutional neural networks[C]//2015 IEEE International Conference on Robotics and Automation. Seattle: IEEE, 2015: 1316-1322.

[5] ten Pas A, Gualtieri M, Saenko K, et al. Grasp pose detection in point clouds[J]. *The International Journal of Robotics Research*, 2017, 36(13/14): 1455-1473.

[6] Murali A, Mousavian A, Eppner C, et al. 6-DOF grasping for target-driven object manipulation in clutter[C]//2020 IEEE International Conference on Robotics and Automation. Paris: IEEE, 2020: 6232-6238.

[7] Qin Y Z, Chen R, Zhu H, et al. S4G: amodal single-view single-shot SE(3) grasp detection in cluttered scenes[C]// Proceedings of the 4th Conference on Robot Learning. London: PMLR, 2020: 53-65.

[8] Liang H Z, Ma X J, Li S, et al. PointNetGPD: detecting grasp configurations from point sets[C]//2019 International Conference on Robotics and Automation. Montreal: IEEE, 2019: 3629-3635.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 6000-6010.

[10] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: hierarchical vision Transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 10012-10022.

[11] Jha D, Smedsrud P H, Riegler M A, et al. ResUnet++: an advanced architecture for medical image segmentation[C]// 2019 IEEE International Symposium on Multimedia. San Diego: IEEE, 2019: 225-230.

[12] Gokcesu K, Gokcesu H. Generalized Huber loss for robust learning and its efficient minimization for a robust statistics[EB/OL]. (2021-08-21)[2026-02-26]. https://arxiv.org/abs/2108.12627.

[13] Choy C, Gwak J, Savarese S. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3075-3084.

[14] Depierre A, Dellandréa E, Chen L M. Jacquard: a large scale dataset for robotic grasp detection[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid: IEEE, 2018: 3511-3516.

[15] Chang A X, Funkhouser T, Guibas L, et al. ShapeNet: an information-rich 3D model repository[EB/OL]. (2015-12-10)[2026-02-26].https://arxiv.org/abs/1512.03012.

[16] Fang H S, Wang C X, Gou M H, et al. GraspNet-1Billion: a large-scale benchmark for general object grasping[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11444-11453.

[17] Morrison D, Corke P, Leitner J. Learning robust, real-time, reactive robotic grasping[J]. *The International Journal of Robotics Research*, 2020, 39(2/3): 183-201.

[18] Zhou X W, Lan X G, Zhang H B, et al. Fully convolutional grasp detection network with oriented anchor box[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid: IEEE, 2018: 7223-7230.

[19] Gariépy A, Ruel J C, Chaib-draa B, et al. GQ-STN: optimizing one-shot grasp detection based on robustness classifier[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau: IEEE, 2019: 3996-4003.

[20] Zhang H B, Lan X G, Bai S T, et al. ROI-based robotic grasp detection for object overlapping scenes[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau: IEEE, 2019: 4768-4775.

[21] Ainetter S, Fraundorfer F. End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB[C]//2021 IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021: 13452-13458.

[22] Kumra S, Kanan C. Robotic grasp detection using deep convolutional neural networks[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver: IEEE, 2017: 769-776.

[23] Wang S C, Zhou Z L, Kan Z. When Transformer meets robotic grasping: exploits context for efficient grasp detection[J]. *IEEE Robotics and Automation Letters*, 2022, 7(3): 8170-8177.

(下转第 151 页)